Homework 8 Solutions

This homework assignment consists of analyzing data from a one-way classification experiment. Six nitrogen-based fertilizers were assigned to plots of land at random. The only way to classify a plot of land is by which nitrogen source it received.

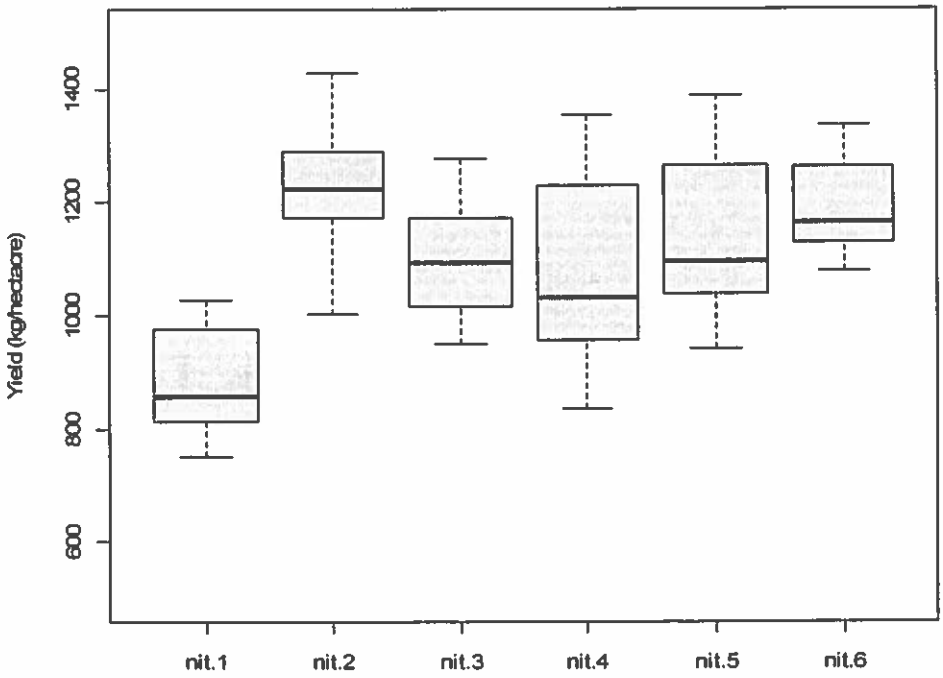
Assumptions: We assume that

(a) The samples of sugar beet yields are independent (plausible because randomization was used to assign nitrogen sources to plots)
(b) Sugar beet yields (Y) are normally distributed in each nitrogen group
(c) The population variance for sugar beet yields is the same for each nitrogen group.

In order to get a first impression of the data (and the plausibility of the second and third assumptions), let's construct boxplots of the samples.

```
# Enter the data
nit.1 = c(814.8,813.2,974.9,862.0,750.8,769.0,1026.0,849.4,946.3,997.9)
nit.2 = c(1235.3,1185.9,1117.0,1171.8,1284.7,1211.5,1288.9,1001.4,1428.4,1373.6)
nit.3 = c(1157.5,1236.1,1074.3,1171.5,1031.3,1015.9,950.1,1108.5,1275.8,999.4)
nit.4 = c(955.0,1039.4,1318.6,926.9,1230.1,835.3,1013.8,1128.3,1023.7,1353.5)
nit.5 = c(1070.0,1153.1,940.1,998.5,1264.3,1351.1,1117.5,1389.3,1037.1,1047.3)
nit.6 = c(1077.2,1137.7,1187.4,1335.8,1262.6,1126.7,1081.6,1134.6,1272.0,1231.3)

# Side by side boxplots
boxplot(nit.1,nit.2,nit.3,nit.4,nit.5,nit.6,
      xlab="",names=c("nit.1","nit.2","nit.3","nit.4","nit.5","nit.6"),
      ylab="Yield (kg/hectacre)",ylim=c(500,1500),col="grey")
```



Initial impression: I am mildly concerned about the constant variance assumption among the six different nitrogen treatments. Specifically, treatment groups 2, 4, and 5 look slightly more variable than nitrogen groups 1, 3, and 6. Remember, that an analysis of variance requires a constant variance assumption among the treatment groups! On the other hand, this may be a bit of an overreaction on my part because the sample sizes here are small (10 plots per nitrogen source), and it is hard to make definite assessments on population level parameters with small samples. Therefore, we proceed but with some minor caution.

Strategy: We will first test the equality of (population) mean yields for the six nitrogen levels. If this test yields a significant result (that is, we reject the equal-population-mean hypothesis), then we will use Tukey confidence intervals to detect where the population mean differences are.

Overall F test: Our first goal is to test the hypothesis of equal treatment population means:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$
versus
$$H_1: \text{at least one } \mu_i \text{ is different.}$$

Here μ_i denotes the population mean yield (kg/hectacre) for the ith nitrogen source (i = 1, 2, ..., 6).

Here is the R code that I used:

```
yield = c(nit.1,nit.2,nit.3,nit.4,nit.5,nit.6)
# Create a treatment indicator variable
nitrogen.type = c(
 rep("nit.1",length(nit.1)),
 rep("nit.2",length(nit.2)),
 rep("nit.3",length(nit.3)),
 rep("nit.4",length(nit.4)),
 rep("nit.5",length(nit.5)),
 rep("nit.6",length(nit.6))
      )

nitrogen.type = factor(nitrogen.type)

data = data.frame(yield,nitrogen.type)
anova(lm(yield ~ nitrogen.type))
```

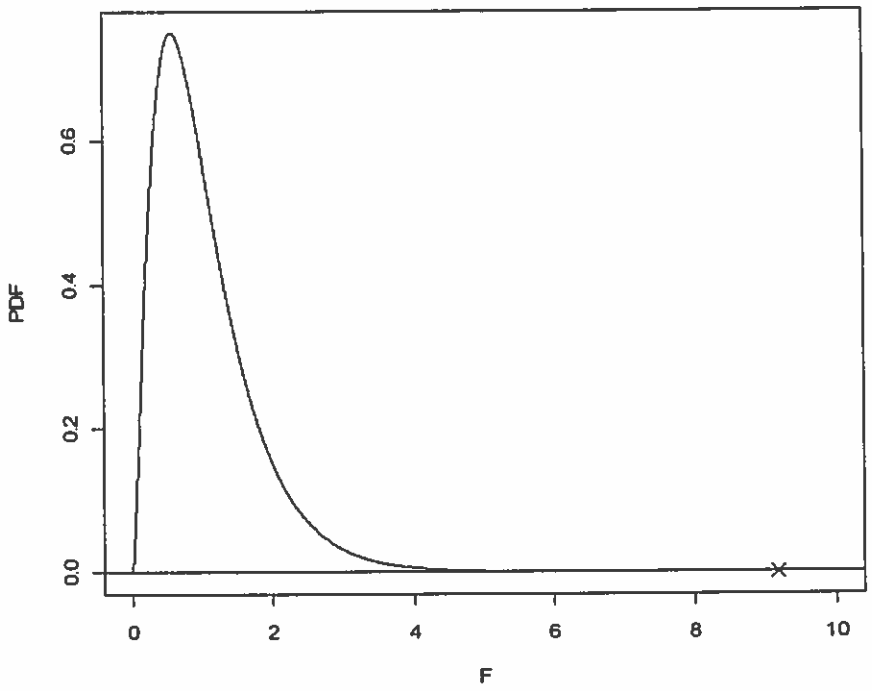
```
> anova(lm(yield ~ nitrogen.type))

Analysis of Variance Table

Response: yield
              Df Sum Sq Mean Sq F value    Pr(>F)
nitrogen.type  5 738684  147737  9.1894 2.262e-06 ***
Residuals     54 868151   16077
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: The F statistic, F = 9.1894, can be used to test the null hypothesis of equal population means (versus the alternative that the means are not equal). Note that if the null hypothesis is true, then the F statistic from the overall ANOVA follows an F(5,54) sampling distribution. Therefore, we can plot this value (F = 9.1894) on this distribution to see if it is "unreasonable" or not. This is done below:

```
# Plot F(5,54) pdf
f = seq(0,10,0.001)
pdf = df(f,5,54)
plot(f,pdf,type="l",lty=1,xlab="F",ylab="PDF",ylim=c(0,0.75))
abline(h=0)
points(x=9.1894,y=0,pch=4,cex=1.5)
```



Interpretation: Clearly, the value F = 9.1894 is not a "reasonable" value from this distribution (p-value = 0.000002262). This suggests strongly that the population mean yields for the different nitrogen sources are different. Note that this conclusion is rather limiting—we don't know where the differences are. We only know (as of now) that at least one population mean yield is different from the others.

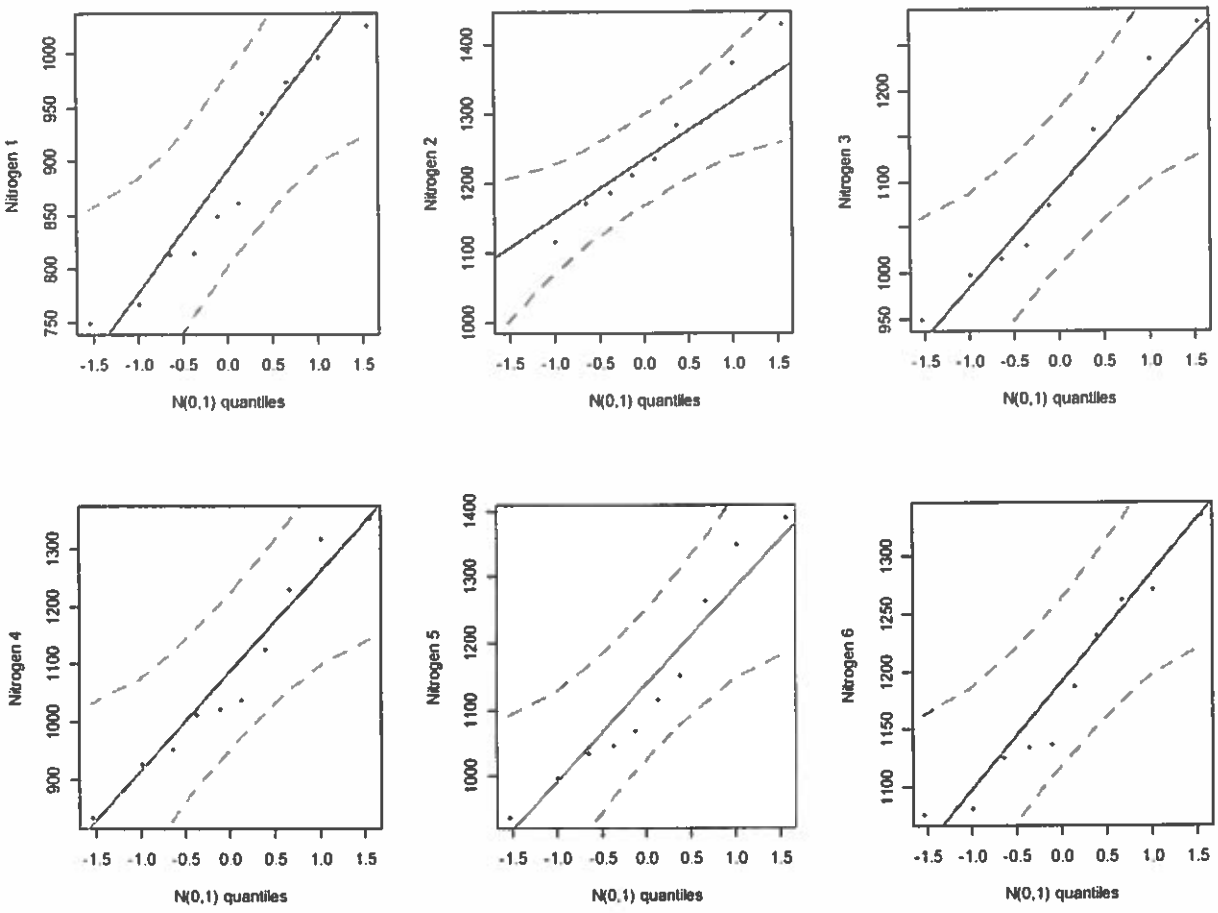
Checking the normality assumption: Remember that our second assumption is that the distribution of sugar beet yields for each nitrogen source is normally distributed. We can diagnose the normality assumption for each group by constructing separate qq plots under the normal assumption.

Of course, it is hard to get clear-cut answers here—we have very small samples (10 yields per nitrogen source). Fortunately, remember that the overall F test (that we just did) is robust to normality departures. In other words, the F test provides reliable inference about the population means even when the normality assumption doesn't hold exactly.

The six qq plots are on the next page:

```
# Load car library
library(car)
# Create a 2 by 3 matrix of plots in R
par(mfrow = c(2,3))
```

```
# qq plots  (under normality)
qqPlot(nit.1,distribution="norm",xlab="N(0,1) quantiles",ylab="Nitrogen 1",pch=16)
qqPlot(nit.2,distribution="norm",xlab="N(0,1) quantiles",ylab="Nitrogen 2",pch=16)
qqPlot(nit.3,distribution="norm",xlab="N(0,1) quantiles",ylab="Nitrogen 3",pch=16)
qqPlot(nit.4,distribution="norm",xlab="N(0,1) quantiles",ylab="Nitrogen 4",pch=16)
qqPlot(nit.5,distribution="norm",xlab="N(0,1) quantiles",ylab="Nitrogen 5",pch=16)
qqPlot(nit.6,distribution="norm",xlab="N(0,1) quantiles",ylab="Nitrogen 6",pch=16)
```



There are no major red flags here, an expected outcome with so little information about each population group.

Follow-up analysis: We now proceed with a Tukey follow-up analysis that constructs (simultaneous) pairwise intervals. I used an overall confidence level of 95 percent.

```
> TukeyHSD(aov(lm(yield ~ nitrogen.type)),conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(yield ~ nitrogen.type))

$nitrogen.type
                diff        lwr       upr     p adj
```

```
nit.2-nit.1  349.42  181.88841 516.95159 0.0000014
nit.3-nit.1  221.61   54.07841 389.14159 0.0033938
nit.4-nit.1  202.03   34.49841 369.56159 0.0095907
nit.5-nit.1  256.40   88.86841 423.93159 0.0004669
nit.6-nit.1  304.26  136.72841 471.79159 0.0000248

nit.3-nit.2 -127.81 -295.34159  39.72159 0.2309852
nit.4-nit.2 -147.39 -314.92159  20.14159 0.1150259
nit.5-nit.2  -93.02 -260.55159  74.51159 0.5761005
nit.6-nit.2  -45.16 -212.69159 122.37159 0.9669432

nit.4-nit.3  -19.58 -187.11159 147.95159 0.9993155
nit.5-nit.3   34.79 -132.74159 202.32159 0.9895906
nit.6-nit.3   82.65  -84.88159 250.18159 0.6920554

nit.5-nit.4   54.37 -113.16159 221.90159 0.9287675
nit.6-nit.4  102.23  -65.30159 269.76159 0.4725848

nit.6-nit.5   47.86 -119.67159 215.39159 0.9577246
```

Interpretation: Note that each pairwise 95 percent interval with nitrogen group 1 (i.e., the control group; no nitrogen) does not include zero and that each interval contains all positive values. This means that, with an overall confidence level of 95 percent, the remaining 5 nitrogen groups produce higher population mean yields than the control group.

On the other hand, we see that all of the other comparisons (i.e., comparing any nitrogen group to any other non-control group) involve confidence intervals which do include zero (equivalently, adjusted p-value greater than 0.05). Rather anticlimactically, this analysis suggests that there are no differences in the population mean yields for the remaining five nitrogen groups.

Insofar as the experimenter's question on "which nitrogen group maximizes population mean yield?" our answer, based on the analysis, would be that it doesn't matter which nitrogen source you use—just as long as it is not the control treatment (with no nitrogen). Our Tukey analysis revealed that there are no differences among the 5 remaining nitrogen sources with respect to population mean yield.

Conclusions: There is a significant difference between the control group and any one of the remaining five nitrogen sources in terms of population mean yield. However, there are no significant differences present among the remaining five nitrogen groups.